Web sémantique et recherche d'informations personnelles

Bruno Charre

DIPLOME D'**É**TUDES **S**UPÉRIEURES
SPÉCIALISÉES
INTELLIGENCE **A**RTIFICIELLE
Université Pierre et Marie Curie
4, place Jussieu
75252 Paris cedex 05

INRIA Rhône-Alpes
Zirst - 655 avenue de l'Europe
38334 Saint Ismier Cedex - France

Rapport de Stage

Web sémantique et recherche d'informations personnelles

Présenté par: Bruno CHARRE
Lieu du Stage: INRIA Rhône-Alpes/EXMO
Responsable de Stage: Jérôme EUZENAT

Mai-Septembre 2002

Remerciements

- Merci à Jérôme Euzenat, mon maître de stage, pour son enthousiasme, sa disponibilité et pour m'avoir aidé à m'organiser.
- Merci à Fabien Triolet, mon collègue de bureau, pour ses remarques pertinentes.
- Et merci à Gaëlle Charre, mon épouse depuis le 27 juillet 2002, de m'avoir soutenu tout au long de mon stage.

Table des matières

Introduction

Mon stage d'une durée de 5 mois s'est déroulé au sein de l'action EXMO de l'INRIA Rhône-Alpes. Les domaines abordés tout au long de ce stage s'inscrivent dans le cadre du web sémantique qui a pour but de faciliter les tâches telles que la recherche d'informations personnelles pour son carnet d'adresse, son agenda, sa bibliographie, etc. J'ai choisi d'effectuer mon stage à l'INRIA (Institut National de Recherche en Informatique et en Automatique) car il me semblait opportun de connaître le milieu de la recherche avant de rentrer dans la vie active et le monde de l'industrie. De plus, le sujet qui m'a été présenté était clair, bien défini et intéressant. En outre, il recoupe un grand nombre de matières étudiées cette année et il concerne un projet qui pourrait constituer l'avenir du Web. Enfin, le cadre : une équipe compétente, des installations puissantes et une bonne ambiance m'ont paru bien adapter pour pouvoir effectuer mon stage dans les meilleures conditions possibles.

Je vous présenterai dans un premier temps la société et l'équipe qui m'ont accueilli. Pour présenter l'INRIA, j'ai synthétisé les informations trouvées sur deux sites qui lui sont dédiés :

- http://www.inria.fr
- http://www.inrialpes.fr

Et pour présenter EXMO où Jérôme Euzenat a été mon seul interlocuteur, j'ai recopié sa propre présentation située à l'adresse : http://www.inrialpes.fr/exmo.html

Puis je restiturai le sujet de stage tel qu'il m'a été proposé. Ce sujet décrit d'abord le contexte c'est-à-dire le web sémantique et la recherche d'information personnelle, puis il fait ressortir les objectifs à atteindre dans la partie travail à effectuer.

Nous reviendrons très largement sur le sujet et les objectifs du stage dans le premier chapitre où j'expliquerai ce que nous avons cherché à faire en élaborant une architecture générale et en expliquant le plan du corps de mon rapport.

Présentation de la société et de l'équipe

L'INRIA

Créé en 1967 à Rocquencourt près de Paris, l'INRIA (Institut National de Recherche en Informatique et en Automatique) est un établissement public à caractère scientifique et technologique (EPST) placé sous la double tutelle du ministère de la recherche et du ministère de l'économie, des finances et de l'industrie. L'INRIA a l'ambition d'être au plan mondial, **un institut de recherche au coeur de la société de l'information.**

Créée en décembre 1992, l'unité de recherche INRIA Rhône-Alpes regroupe plus de 400 personnes réparties sur trois sites : la Zirst de Meylan-Montbonnot, le campus universitaire de Grenoble et le site technopolitain de Lyon (dont Lyon-Gerland et le domaine scientifique de la Doua). Dans le contexte de l'INRIA Rhône-Alpes, les activités de recherche se traduisent en quatre pôles de recherche prioritaires notamment «**Aider à la conception et à la création : bases de connaissances, documents multimédia, modèles cognitifs**».

Intégrant des connaissances de plus en plus évoluées, la machine est à même d'assister l'homme dans la réalisation d'activités individuelles ou collectives, qu'il s'agisse de concevoir des documents multimédia ou de constituer des bases de connaissances scientifiques ou techniques. L'équipe EXMO qui s'intéresse aux «**Echanges de connaissance structurée médiatisés par ordinateur**» s'inscrit parfaitement dans ce pôle de recherche.

EXMO

EXMO étudie l'échange de connaissance structurée et formalisée. La connaissance est représentée dans des langages formellement définis. Ils peuvent aller de XML - métalangage structuré mais sans sémantique - aux langages de représentation de connaissance - structurés, sémantiquement définis mais spécialisés.

Le but de l'action EXMO est le développement d'outils théoriques et logiciels pour aider à l'organisation, la manipulation, la composition et la présentation d'éléments de connaissance structurés lors de la communication entre humains. Dans le processus de communication, l'ordinateur peut introduire une plus-value à son rôle de médium et de mémoire en accomplissant des tâches comme le formatage, le filtrage, la catégorisation, le test de consistance ou la généralisation.

Assurer l'adéquation et l'intelligibilité de la connaissance pour les interlocuteurs nécessite le développement d'une compréhension abstraite des re-

présentations et des transformations qui leur sont appliquées. Les travaux de l'action EXMO sont focalisés sur deux aspects. L'aspect transformation rend compte des modifications de la connaissance pendant la communication alors que l'aspect communication concerne la préservation de l'intelligibilité de la connaissance transformée.

Présentation du sujet du stage

Web sémantique et recherche d'information personnelle

Le web sémantique est décrit généralement comme un web destiné aux machines. Disposer d'un web dont le contenu est abordable par les machines peut apporter divers bénéfices :

- L'automatisation de nombreuses tâches fondées sur le contenu comme la recherche de ressources ayant un contenu particulier, la comparaison du contenu de ressources (pages, bases de données, ontologies, etc.). Le web sémantique permettrait de résoudre la relative difficulté de trouver de l'information sur le web.
- L'automatisation de tâches liées à la mise en relation de ce contenu, comme l'articulation de réponses fournies par plusieurs ressources.
- La description de ressources informatiques (services) par leurs conditions d'activations, résultat, qualité, etc. permet d'imaginer la recherche, l'invocation et la connexion automatique de ces ressources.

Notre conviction est que les machines doivent aider les utilisateurs dans leurs micro-tâches de recherche d'informations personnelles autant que dans les macro-tâches de synchronisation entre entreprises.

Travail escompté

Le but du stage est de développer des outils pour aider les utilisateurs à rechercher de l'information personnelle (il s'agit ici d'information de type agenda et carnet d'adresse) sur le web et à l'intégrer dans des formats structurés et standards. Le contenu du stage a deux optiques :

- technologique : il s'agit de créer une bibliothèque d'importation, d'exportation et de manipulation du format vCard/vCal en XML et RDF. Ce format est utilisé par la plupart des outils de gestion d'information personnels (Palm, PocketPC, Netscape, Eudora, calendar, Outlook, AddressBook). Cette bibliothèque sera intégrée à l'environnement de développement de transformation XML Transmorpher. La programmation se fait en Java en utilisant SAX et XSLT.

- applicative : il s'agit de développer une application permettant de rechercher de l'information complétant un ou plusieurs champs de vCard et d'en donner une mesure de confiance. La recherche pourra être réalisée sur le web en utilisant des techniques de wrappers ou en s'appuyant sur une base de vCard existantes. Il faudra développer un modèle de propagation de la confiance intégrant la confiance dans les sources, les dates de création et les préférences de l'utilisateur.

Chapitre 1

Architecture générale

Le web sémantique, qui n'est pas destiné à remplacer l'actuel mais à l'améliorer, permettrait de rendre les outils de recherches plus intelligents et plus faciles à utiliser. Le sens contenu dans l'information donnée au moteur de recherche sera mieux défini et nous pourrons alors utiliser notre ordinateur de manière plus efficace. Une plus longue explication du web sémantique sera donnée en annexe.

Le problème est que ce projet rend sceptique. Pour combattre l'incrédulité, il faut donc créer quelques outils permettant de prouver la faisabilité et l'intérêt d'une telle approche pour convaincre à la fois le public et les professionnels.

1.1 Analyse globale

Dans cette optique, nous avons donc essayé d'élaborer des outils permettant de produire, pour une communauté, des informations extraites du "web sémantique", dans le format qu'elle utilise couramment (vCal/vCard, Palm, HTML, PDF...). L'idée est de construire un système capable de rassembler et d'exploiter de l'information sur le web concernant les évènements et les personnes. Plus spécifiquement, il s'agit de partir d'un format de données de type vCal et vCard exprimé en XML(Extensible Markup Language) ou en RDF(Resource Description Framework). Puis, il faut faire de l'inférence de ces données (à l'aide d'heuristiques à développer, par exemple si quelqu'un travaille à un endroit, si je trouve une personne avec le même nom dans l'annuaire habitant une commune limitrophe de un endroit alors j'ai son numéro de téléphone) et de leur appliquer des contraintes (test de consistance principalement). Enfin, nous devons être capables d'attribuer et de propager une mesure de confiance à partir des informations hétéroclites (le site d'une

Université est plus fiable que la vCard d'une personne pour me dire s'il y travaille).

1.2 Méthodologie

Les besoins de rassembler et de communiquer de l'information électronique augmentent avec entre autres l'avénement des ordinateurs de poche. L'échange de données personnelles se produit chaque fois que deux individus ou plus communiquent. De tels échanges incluent fréquemment l'échange d'information, telle que des cartes de visite professionnelle, des numéros de téléphone, des adresses, des dates, des rendez-vous, des références bibliographiques, etc. L'électronique et les télécommunications peuvent aider à s'assurer que l'information est rapidement et sûrement communiquée, stockée, organisée et facilement localisable quand c'est nécessaire en utilisant des formats de représentation de la connaissance. Pour pouvoir nous y retrouver et choisir le format répondant à nos attentes, nous allons dans un premier temps étudier les formats de PIM(Personal Information Management) existants. Le but étant d'obtenir à terme une cohérence et une liaison entre chaque sorte de données.

Nous avons axé nos recherches sur les deux formats créés par le consortium versit qui sont vCard , une carte de visite professionnelle électronique, et vCalendar , un format d'échange d'agendas et de planing électronique. Nous avons choisi ces formats car ils sont utilisés par la plupart des outils de gestion d'informations personnelles. Une fois la représentation connue, il faut pouvoir la manipuler. Pour cette raison, nous avons choisi de transformer ces formats en XML(eXtensible Markup Language) qui est portable, standard et facile à manipuler grâce à des API Java tel que SAX ou DOM. Ainsi, j'ai créé une bibliothèque d'importation et d'exportation du format que nous avons intégré à l'environnement de développement de transformation XML Transmorpher. Transmorpher est un logiciel développé par l'action Exmo. C'est un système de spécification et d'exécution de flux de transformations. Cette bibliothèque créée m'a servi pour ouvrir une vCard existante dans mon application de recherche d'informations personnelles. J'expliquerai les modifications à effectuer pour pouvoir manipuler une vCard dans mon application.

Dans la dernière partie, nous avons choisi de nous focaliser sur la description de personne avec le format vCard. Lorsque les fichiers vCard peuvent être manipulés, il faut pouvoir trouver puis traiter l'information cherchée. Donc j'énumérerai dans un premier temps, les sources de données permettant d'obtenir des données. Puis j'analyserai plus en détails les dépendances

entre attributs pour savoir à partir de quels champs je peux trouver un autre champ. Je suggérerai ensuite un "modèle de certitude" qui restera néanmoins très simple. Et enfin, je présenterai la réalisation dans sa globalité en décrivant les expérimentations ainsi que les résultats obtenus.

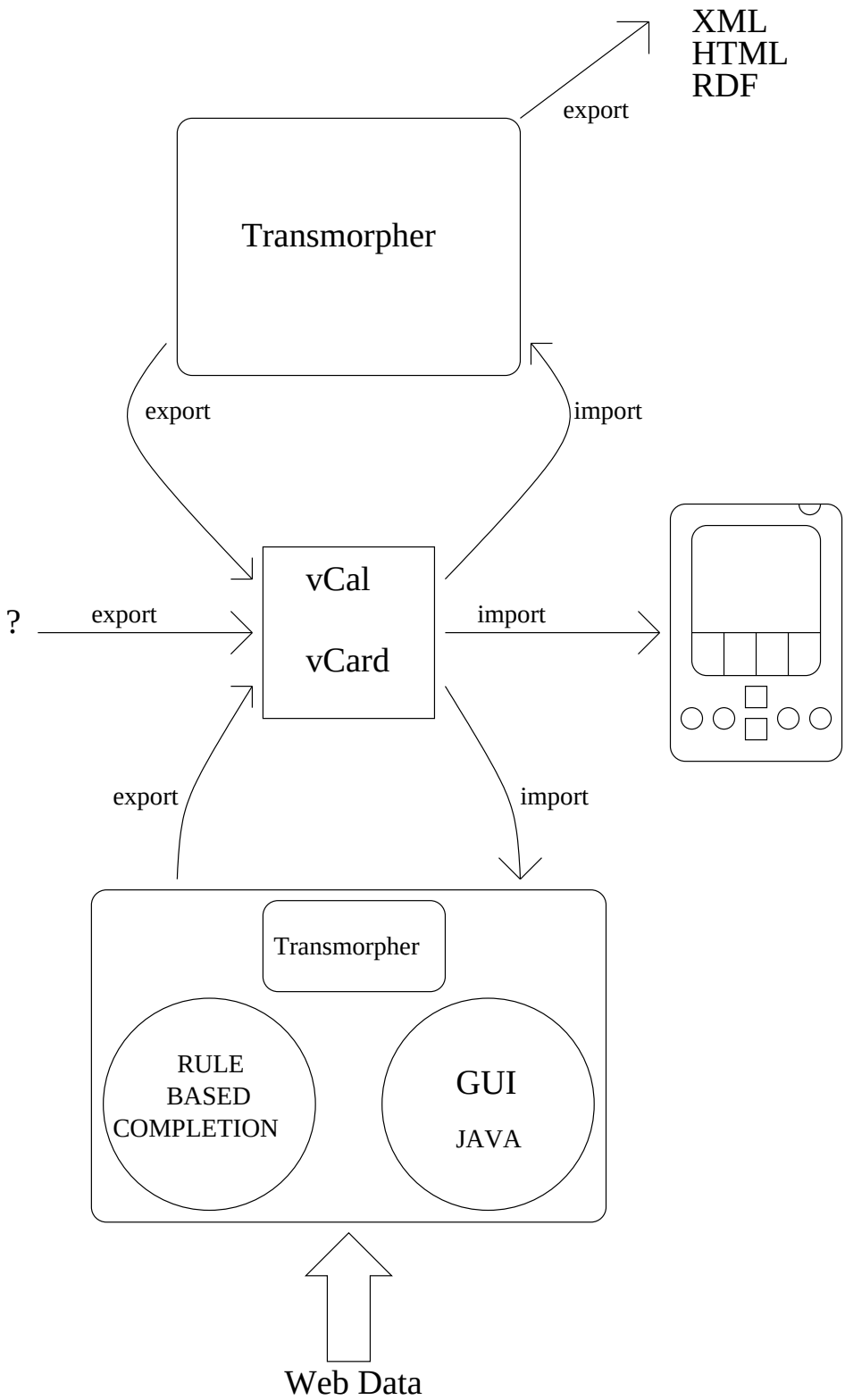

FIG. 1.1 – Architecture générale

Chapitre 2

Formats de PIM

Les ordinateurs traiteront et délivreront des informations beaucoup plus précises si les données sur le Net et leurs relations aux autres données sont plus précises. Il existe un grand nombre de données structurées par des formats connus. Dans la première partie nous nous intéresserons aux formats représentant des personnes. Puis nous verrons ceux spécifiques au calendrier. Ensuite, nous survolerons les formats adaptés aux documents. Enfin, nous essaierons de mettre toutes ces informations en relation grâce à un schéma indiquant les dépendances entre chaque format.

2.1 Personnes

Il existe plusieurs formats de données pour représenter une personne. Nous en passerons 3 en revue : les vCard(Virtual Card), un format plus récent FOAF(Friend-Of-A-Friend) et l'annuaire LDAP(Lightweight Directory Access Protocol). Nous les présenterons briévement ici sachant que l'annexe B en synthétise leurs attributs.

2.1.1 vCard

Les vCard[2] existent en plusieurs versions dont la version 3.0 qui a donné naissance à la norme RFC 2426. La version précédente (version 2.1) est encore utilisée par la plupart des logiciels. Il existe plusieurs différences entre ces 2 versions qui sont la création de nouveaux champs, la suppression de certaines fonctionnalités (par exemple l'encodage a été supprimé). Dans l'ensemble je trouve que la version 3.0 est plus structurée en plus du fait qu'elle soit normalisée. La plupart des travaux ont été effectués en s'appuyant sur cette norme. Le plus intéressant de ces projets est la représentation des ob-

jets vCard en RDF/XML[7] qui est une note du W3C. Le RDF/XML est un langage de description des données qui a été conçu pour faciliter le traitement automatique des ressources Web. Il faut savoir qu'une vCard n'est pas spécifique à la représentation d'une personne mais qu'elle a été conçue à l'origine pour représenter d'autres entités telles que les places ou les organisations.

2.1.2 FOAF

FOAF(Friend-Of-A-Friend)[1] a été conçu par un groupe de développeurs indépendants pour illustrer les possibilités de RDF. Il est employé dans un certain nombre de projets dédiés au Web sémantique.

L'idée fondamentale de FOAF est de permettre aux machines de se servir de l'information stockée dans les formats FOAF. Si ces fichiers contiennent des liens vers d'autres documents Web alors, les programmes seront capables de parcourir le Web stockant l'information qu'ils trouvent, gardant une liste de pointeur vers d'autres documents, vérifiant les signatures numériques et construisant des pages Web ainsi que des services de questions/réponses basés sur les documents visités.

Les fichiers FOAF sont justes des documents textes. Ils sont écrits dans la syntaxe XML, et adoptent les conventions RDF, annoté avec DAML. De plus, le vocabulaire FOAF définit quelques constructions utiles qui peuvent apparaître dans les fichiers FOAF, à côté d'autres vocabulaires RDF définis ailleurs. Par exemple, FOAF définit des catégories telles que : 'Person', 'Document ', 'Image ', avec quelques propriétés maniables de ces champs, telles que 'name', 'mbox', 'homepage' etc., ainsi que des liens entre les membres de ces catégories. Par exemple, un type de lien intéressant est 'foaf :depiction'. qui relie une catégorie (par exemple une personne) à une image. Les outils pour FOAF sont basés sur des logiciels qui analysent les documents RDF et se servent de ces propriétés.

Le contenu spécifique du vocabulaire FOAF est détaillé en annexe. Un dispositif intéressant d'un fichier FOAF est qu'il peut contenir des pointeurs vers d'autres fichiers FOAF. Ceci fournit une base pour les outils de recherches pour parcourir un ensemble de fichiers liés entre eux, et se renseigner sur de nouvelles personnes, documents, services, données...

2.1.3 LDAP

LDAP est le protocole d'annuaire sur TCP/IP. Les annuaires permettent de partager des bases d'informations sur le réseau interne ou externe. Ces bases peuvent contenir toute sorte d'informations que ce soit des coordonnées de personnes ou des données systèmes.

LDAP fournit :

- le protocole permettant d'accéder à l'information contenue dans l'annuaire,
- un modèle d'information définissant le type de données contenues dans l'annuaire,
- un modèle de nommage définissant comment l'information est organisée et référencée,
- un modèle fonctionnel qui définit comment on accède à l'information,
- un modèle de sécurité qui définit comment données et accès sont protégés,
- un modèle de duplication qui définit comment la base est répartie entre serveurs,
- des APIs pour développer des applications clientes,
- LDIF, un format d'échange de données.

Les données LDAP sont structurées dans une arborescence hiérarchique qu'on peut comparer au système de fichier Unix. Chaque nœud de l'arbre correspond à une entrée de l'annuaire.

2.1.4 Conclusion

LDAP est un outil où chaque utilisateur décrit ses propres champs. Or, il nous fallait un format plus structuré qui permette de donner un sens à chaque information décrite dans la structure. Pour cette raison, nous aurions très bien pu choisir le format FOAF, mais il est très récent et peu d'applications de PIM l'utilisent. Cependant, FOAF est un format tourné vers l'avenir et je pense qu'il faut faire attention à son évolution. Nous avons finalement choisi le format vCard qui est pris en compte par la plupart des outils de gestion des informations personnelles et qui en plus a donné naissance à une norme. En plus, le milieu du web sémantique s'intéresse à ce format car il a écrit une note pour pouvoir le transformer en RDF.

2.2 Calendrier (event)

Nous allons étudier deux formats pour stocker des agendas ou des plannings. Nous étudierons dans un premier temps vCalendar. Puis nous préciserons les différences avec iCalendar qui est basé sur les travaux de vCalendar.

2.2.1 vCalendar

vCalendar est un format simple permettant de partager des données relatives à la gestion d'agendas et à la planification. Le but de vCalendar est de structurer l'information contenue dans un agenda pour pouvoir échanger des données, des images ou des sons concernant un événement ou une tâche. Pour pouvoir réaliser cet échange, les informations sur la date et l'heure doivent être comprises dans toutes les régions du monde ; même ceux qui emploient différents formats de date, d'heure, et de fuseau horaire. Pour cette raison, le format vCalendar se sert d'un certain nombre de normes existantes.

Ainsi la norme ISO 8601 est une norme internationale pour représenter la date et l'heure. Le format utilisé dans les spécifications de vCalendar pour identifier des langues et des jeux de caractères est basé sur les normes ISO et IETF. La syntaxe utilisée dans les spécifications vCalendar est basée sur la grammaire définie par les normes internet STMP et MIME. L'utilisation de cette syntaxe facilite le transfert des objets vCalendar dans les services de messagerie électronique basés sur MIME.

Le format vCalendar est basé sur deux entités qui sont EVENT et TODO. Un EVENT représente une quantité de temps indiquée sur un calendrier. Par exemple, ce peut être une activité, telle qu'une soutenance d'une heure de 9H00 à 10H00, le vendredi 27 septembre. Un TODO représente une action à réaliser ou une tâche. Par exemple, ce peut être un travail à accomplir comme "rédiger le rapport de stage pour le 9 septembre"

Une grande variété de produits permettent d'importer et d'exporter le format vCalendar. vCalendar peut être employé pour transporter un agenda entre différentes applications tels que les gestionnaires d'informations personnelles, les messageries ou les navigateurs Web.

2.2.2 iCalendar

iCalendar[3] étend en réalité vCalendar et on peut considérer que iCalendar est la version 2.0 de vCalendar. iCalendar se compose de trois RFCs :

- RFC 2445 , Internet Calendaring and Scheduling Core Object Specification (iCalendar) : les spécifications d'iCalendar.
- RFC 2446 , iCalendar Transport-Independent Interoperability Protocol (iTIP)
- RFC 2447 , iCalendar Message-based Interoperability Protocol (iMIP)

Quelques composants ont été rajoutés en plus des EVENT et des TODO :

- VJOURNAL qui fournit des propriétés décrivant une écriture comptable ;
- VFREEBUSY qui permet de signaler si on est libre ou occupé ;

- VTIMEZONE qui fournit un ensemble d'indications sur le fuseau horaire ;
- VALARM qui fournit un ensemble de propriétés pour décrire une alarme.

2.2.3 Conclusion

Les règles de récurrences ont changé mais les applications qui supportent iCalendar devraient supporter les vCalendar. Nous pouvons considérer qu'il existe une version 3.0 xCal (XML iCalendar) voulue par le W3C qui veut revoir et changer le standard iCalendar pour qu'il soit conforme à XML mais ce projet est seulement une ébauche. Nous avons choisi les vCalendar mais la bibliothèque d'importion et d'exportation fonctionne aussi pour les iCalendar.

2.3 Références...Documents

Les formats représentant une bibliographie sont très intéressants notament lors des conférences où un grand nombre de documents doit être référencié. Nous étudierons deux formats appropriés qui sont BibTEX et Dublin Core.

2.3.1 BibTEX

BibTEX[4] est un utilitaire qui gère une bibliographie : il sait extraire, trier et mettre en page des éléments d'une base de données, afin d'en faire un résultat compilable par LATEX. Les données sont contenues dans le fichier .bib où se trouvent les renseignements bibliographiques ainsi que les éléments. Ce fichier contient plusieurs entrées de la forme :

```
type_d_entrée{clef_interne,
              champ_1 = "valeur_1",
              champ_2 = "valeur_2",
              ...
              champ_n = "valeur_n"
}
```

Pour la mise en forme des données BibTEX consulte le fichier de style. Le rôle de ce fichier est d'indiquer comment imprimer (formater) les références. Le rôle des champs sera décrit en annexe ainsi que les différents types d'entrées.

Il existe deux entrées spéciales **@string** et **@preamble**. L'entrée **@string** permet de définir des abréviations et **@preamble** permet d'insérer des commandes (ou du texte) dans le fichier créé par BibTEX.

Un champ intéressant est **crossref** qui permet de faire des références à l'intérieur de la bibliographie. De cette façon on peut citer un document et une partie de ce document, et dans ce cas avoir une référence croisée dans la bibliographie, ou faire hériter à une entrée de la bibliographie les champs d'une autre entrée.

2.3.2 Dublin Core

La norme de métadonnées du Dublin Core[8] est un ensemble d'éléments simples mais efficaces pour décrire une grande variété de ressources en réseau. La norme du Dublin Core comprend 15 éléments dont la sémantique a été établie par un consensus international de professionnels provenant de diverses disciplines telles que la biblioéconomie, l'informatique, le balisage de textes, la communauté muséologique et d'autres domaines connexes.

On trouvera une description de l'ensemble des éléments du Dublin Core en annexe. Chaque élément est optionnel et peut être répété. Chaque élément possède également un ensemble limité de qualificatifs, des attributs qui peuvent être utilisés afin de raffiner davantage (et non pas étendre) la signification de l'élément. L'Initiative de métadonnées du Dublin Core (IMDC) a défini, en juillet 2000, des façons normalisées de "qualifier" les éléments au moyen de différents types de qualificatifs. Un registre de qualificatifs conformes aux "meilleures pratiques" de l'IMDC est en cours de construction.

Bien que le Dublin Core favorise la description d'objets ressemblant à des documents (car la description des ressources textuelles traditionnelles est une activité bien maîtrisée), son usage pour la description des ressources ne ressemblant pas à des documents traditionnels va dépendre, jusqu'à un certain point, des similitudes entre les métadonnées de ces nouveaux documents par rapport aux métadonnées habituelles d'un document. Il va aussi dépendre des objectifs visés par les métadonnées de ces nouveaux documents.

Dublin Core a pour objectif de concilier les caractéristiques suivantes :

- simplicité de création et de gestion,
- sémantique communément comprise,
- envergure internationale,
- extensibilité.

2.4 Conclusion

Nous avons vu différents formats dont nous avons isolé vCard et vCalendar pour nôtre travail. Une première remarque à ce stade est que ces formats sont dévelppés indépendemment même si souvent ils pourraient partager des éléments. Ainsi, le schéma suivant présente la relation entre ces données.

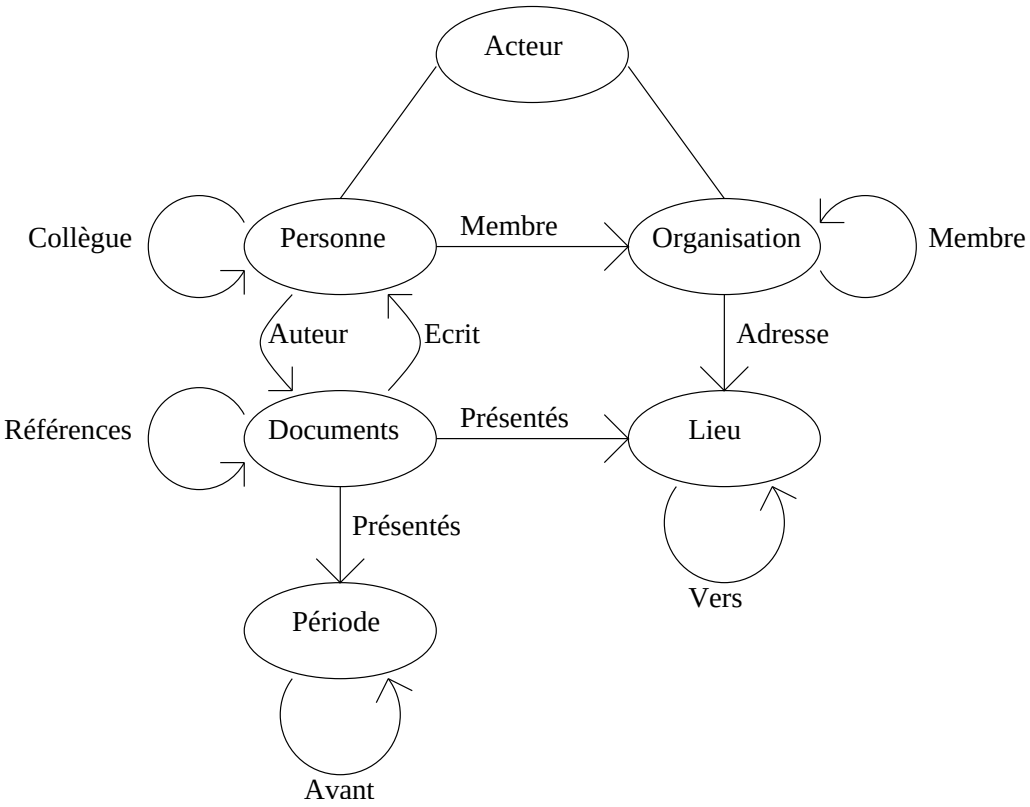

FIG. 2.1 – Liaison entre les connaissances

Chapitre 3

Manipulations des informations de PIM

Pour effectuer ce travail, j'ai dû me pencher sur la façon de manipuler les Linformations PIM.Le but était de transformer les formats vcs dans un format XML que nous maîtrisons beaucoup mieux. Pour atteindre ce but, je me suis servi uniquement de JAVA et des outils mis à ma disposition pour pouvoir traiter des fichiers XML.

Cela m'a permis de découvrir que plusieurs méthodes permettent de parser du XML [6] comme SAX(Simple Api for XML) et DOM(Document Object Model). DOM et SAX ne remplissent pas les mêmes objectifs. DOM permet de construire un arbre représentant le document XML et offre une API pour les manipuler tandis que SAX envoie des messages chaque fois qu'elle rencontre un élément particulier dans le fichier XML et des handlers sont capables d'interprêter ces messages, et d'agir en conséquence. Donc l'API SAX est plus rapide que l'API DOM pour gérer les événements. Mais, au contraire de SAX, DOM est une structure de données où l'on peut stocker de l'information.

J'ai découvert aussi quelques outils qui permettent de piloter du XSLT[5] comme Transmorpher développé par l'action EXMO. Transmorpher est un environnement permettant de définir et d'exécuter des flux de transformations complexes sur des documents XML. Il s'appuie en partie sur le langage de transformation XSLT (XML Stylesheet Language - Transformation) .Il permet de décrire l'enchaînement complexe de transformations simples impliquant la gestion de plusieurs documents simultanément. C'est un langage de transformation s'exprimant en XML, portable, ouvert vers d'autres moteurs de transformation et cependant suffisant pour décrire des systèmes de transformations complexes.

3.1 Import

Quand j'ai débuté mon stage, il existait déjà un outil permettant de transformer une vCal en XML programmé en Perl par Dan Connolly. J'ai alors étudié le fonctionnement de ce programme. Je voulais surtout élaborer un format de sortie en XML qui soit cohérent. Je me suis aperçu assez vite que ce programme effectuait une transformation simple qui ne faisait pas ressortir toutes les informations qui nous étaient utiles tels que les éléments des champs composés. Effectivement, il est important de bien structurer le format vCal/vCard en XML. Par exemple, lors de notre recherche d'informations avec vCard, nous aurons besoin de connaître le nom de famille d'une personne pour pouvoir trouver son numéro de téléphone. Pour cette raison, il faut analyser la valeur du champ composé N pour pouvoir sectionner l'information en données plus précises. Ceci dit, le programme Perl m'a permis d'avoir une notion de la grammaire utilisée par le format vCalendar.

Il a fallu cependant que j'étudie en détails la grammaire formelle de vCard et vCalendar pour pouvoir analyser les fichiers et les transformer en XML. De plus, il fallait sortir une structure qui soit à peu prés similaire pour les vCard et les vCalendar pour pouvoir avoir une seule fonction d'importation. La grammaire formelle utilisée est la suivante :

```
[^;:]+ (; ([^=;:]+ =)? [^=;:]+ )* : value.
```

De cette façon, nous avons pu analyser notre fichier en sachant que pour chaque champs BEGIN il y a un END correspondant et que seules les caractéristiques connues commencent en début de ligne. Nous avons dû prévoir des traitements appropriés aux champs composés ainsi que les champs encodés pour lesquels nous nous sommes servis de la bibliothèque JAVA Mail.

Puis, nous avons dû prendre une décision concernant le format de sortie en XML. Il existe un certain nombre de DTD permettant d'exprimer les vCard et les vCalendar en XML mais elles ne définissent que les dernières versions de ces formats. Je me suis basé sur le format de sortie du programme Perl qui fait correspondre à un champ vcs une balise XML en ajoutant des balises pour les valeurs composées tel que N, ADR, etc... Les attributs des fichiers vcs correspondent aux attributs des balises XML. Certains attributs sont définis seulement par leur valeur sans leur nom qui est facultatif. Dans ce cas là, il faut déterminer le nom de l'attribut car les attributs dans les fichiers XML sont composés obligatoirement d'un nom et d'une valeur.

Dans un premier temps, j'ai écrit un programme qui analysait un fichier vcs et qui l'imprimait en sortie dans un fichier XML. Transmorpher se ser-

vant de l'API SAX, il fallait que je génère des événements SAX pour que l'application puisse à son tour traiter et retransmettre ces événements. Lors de l'insertion dans Transmorpher, nous avons analysé comment intégrer ma bibliothèque dans le logiciel en respectant la logique des autres classes. La conclusion était que ma classe devait implémenter l'interface XMLReader, comme le font les autres parser comme Xerces, même si le fichier d'entrée n'est pas du XML.

3.2 Export

Un de mes premiers travaux consistait à travailler sur l'interconnexion de programmes de conférences avec agendas et calendriers. Le but de ce travail consistait à exporter un fichier représentant un programme de conférence en différents formats : HTML, vCal, RDF/iCal, DAML+OIL. Cette première approche m'a permis de me familiariser avec l'outil Transmorpher qui permet de piloter du XSLT depuis JAVA et avec XSL pour l'écriture des feuilles de style. Grâce à Transmorpher, nous pouvons dispatcher le fichier d'entrée et appliquer une feuille de style à chaque sortie du dispatcher pour obtenir les différentes sorties voulues. Et j'ai pu au passage créer une DTD d'après le fichier XML représentant le programme de la conférence de l'ISCW 2002 (1st International Semantic Web Conference) qui a pu être réutilisée lors de l'ECAI 2002 (15th European Conference on Artificial Intelligence).

Au tout début de ce mini-projet, nous transformions directement le programme de conférence en XML en objet vCal grâce à une feuille de style XSL. Or nous voulions obtenir un format intermédiaire représentant le fichier vCal en XML. J'ai donc créé une feuille de style XSL pour obtenir le format désiré et une autre pour passer de ce format au vrai format vCal.

Le problème d'une telle manœuvre est qu'avec une simple feuille XSL nous ne pouvons pas encoder les parties qui doivent l'être. Dans notre exemple, nous n'avions pas ce problème car le seul encodage utilisé pour décrire les conférences est QUOTED-PRINTABLE. Or ce codage de données est facile à implanter bien que des différences subsistent entre les différents gestionnaires d'informations personnelles. Mais, il n'en est pas de même avec toutes les sortes d'encodage permises par vCal et vCard. C'est pour cette raison que nous avons dû réaliser l'exportation en JAVA.

La classe qui permet d'écrire des données dans un format vcs doit traiter les événements reçus du Transmorpher. Pour faire cela, notre classe implémente l'interface ContentHandler de SAX qui permet de gérer les éléments et les valeurs associées à ces éléments.

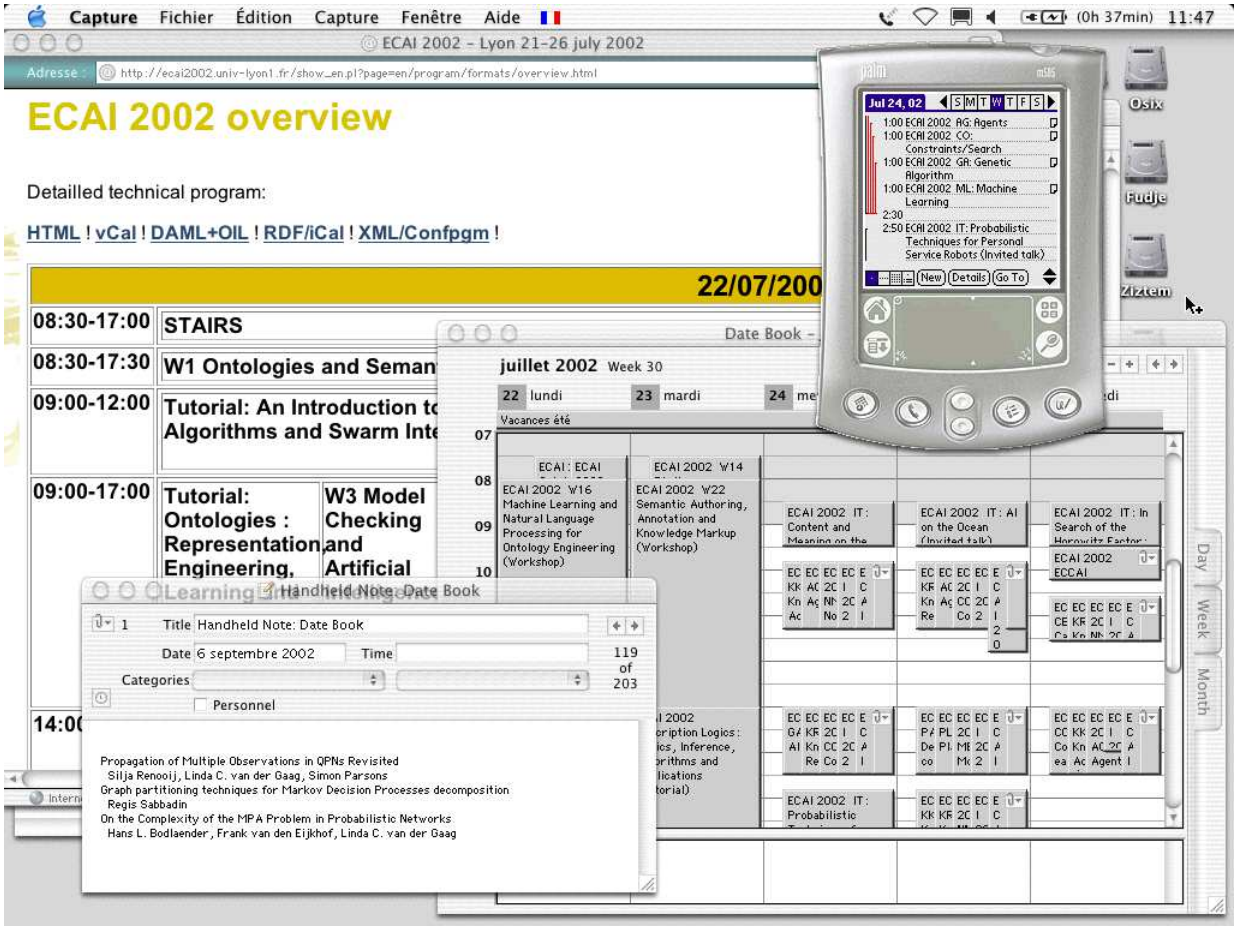

FIG. 3.1 – Exemple des différents formats de sortie dans différents outils : Browser HTML, Palm Destop, Palm OS.

3.3 VCardGUI

Une fois la bibliothèque d'importation et d'exportation du format vCal/vCard intégrée à l'environnement de développement de transformation XML Transmorpher, nous devions nous en servir dans l'application VCardGUI. Ainsi nous aurions pu exporter notre fichier vCard dans plusieurs formats de sortie. Le but premier étant d'aller le plus loin possible dans la recheche d'informations, nous n'avons pas exploité cette idée. De plus, Transmorpher utilise des événements SAX qui ne sont pas adaptés pour garder en mémoire les données.

Pour l'instant, le bouton IMPORT permet d'importer les fichiers vCard dans l'interface graphique et le bouton EXPORT permet d'exporter les don-

nées contenues dans VCardGUI dans un seul format : vCard.

Le problème à résoudre était d'insérer l'arbre DOM dans mon application car il fallait que la valeur du nœud soit écrites dans le bon TextField. Au début, je pensais donner des noms compréhensifs à mes TextField et comparer la valeur de mon noeud avec le nom du TextField. Mais ce n'était pas une solution idéale car nous devions transformer la chaîne de caractères représentant le champ traité en l'objet TextField correspondant. La solution a donc été de créer une table de hachage de TextField.

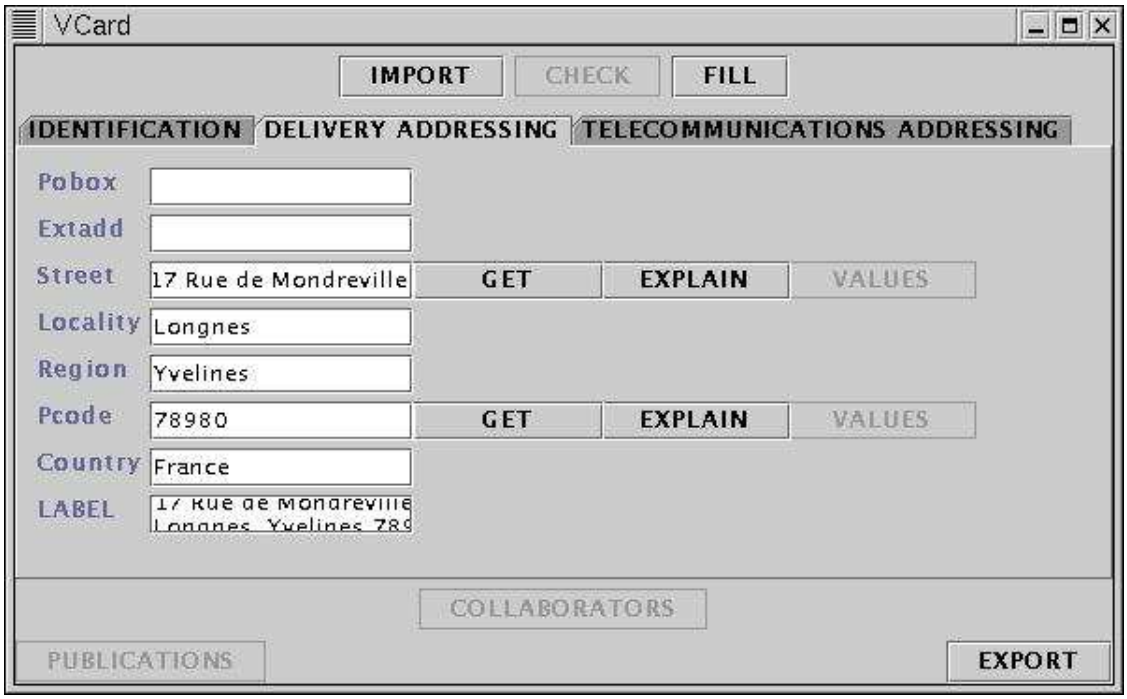

FIG. 3.2 – VCARDGUI après l'importation de données

Chapitre 4

Source d'information

Le plus grand problème lors d'une recherche sur le Web n'est pas de ne rien trouver, mais au contraire d'obtenir un trop grand nombre de résultats qu'il est impossible de trier.

Il faut donc savoir où et comment chercher (pertinence des données sur Internet). Pour résoudre ce problème, nous expliquerons où nous cherchons l'inforamtion, quelles sont les dépendances entre attributs pour pouvoir compléter notre requête, et comment nous avons intégré ces données dans notre application.

4.1 Les sources de données permettant d'obtenir des données

Il y a essentiellement deux lieux où nous sommes susceptibles de trouver de l'information pour remplir les champs d'une vCard. Le lieu auquel nous pensons au départ est bien évidemment le Web qui représente une source de données intarissable. Mais il aurait aussi été possible de s'appuyer sur une base de vCard existantes en stockant l'information dans une base de données et ainsi chercher l'information voulue grâce à une requête spécifique. Je me suis limité à utiliser le Web pour une question de temps. Il est vrai que l'utilisation d'une base de vCard aurait certainement pu nous apporter des réponses plus intéressantes pour certains champs des vCard. Par exemple si nous avons un ensemble de vCard représentant des personnes travaillant au même endroit, il est fortement probable que l'adresse d'expédition soit la même pour toutes les autres personnes travaillant à cet endroit.

Le web est une source d'informations gigantesque et nous voulons obtenir une information précise. Pour cette raison, il faut savoir où trouver l'information et chercher au bon endroit.

Par exemple,pour trouver le téléphone nous devons chercher dans l'annuaire et faire la requête suivante :
http://www.pagesjaunes.fr/pb.cgi?faire=decode_input_image&DEFAULT_ACTION=
bf_inscriptions_req&mode=web&SESSION_ID=GD-49373D0-FD42&VID=FF-4741480-42CE6
FR&srv=PB&TYPE_RECHERCHE=ZZZ&input_image=&FRM_NOM=$Nom&FRM_
LOCALITE=$Localite

4.2 Les dépendances entre attributs

Pour pouvoir faire une recherche fructueuse, il faut savoir quelles informations sont nécessaires pour l'obtention d'une autre information. Je vais donc mentionner pour quelques champs de la vCard décrit dans les annexes, les données obligatoires et celles qui sont facultatives.

Champs à trouver	Champs obligatoires	Champs facultatifs
PHOTO	Family	Given, Orgname
Given	Family, Locality, Country	Street
Street	Family, Locality, Country	Given
Téléphone	Family, Locality, Country	Given, Street
TZ	Locality, Country	
URL	Family	Given, Orgname

4.3 Un "modèle de certitude"

Le modèle de certitude élaboré doit être simple. Il s'appuie sur plusieurs facteurs qui sont :

- la multiplicité des réponses : il paraît évident que si nous obtenons plusieurs réponses différentes pour une même requête, la certitude que la réponse fournie est la bonne s'amenuise.
- la date de la dernière mise à jour : nous pouvons penser que les pages les plus récentes ont des informations plus proches de la vérité que les plus anciennes.
- la complétion. Entre EUZENAT J. et EUZENAT Jérôme on préférera la deuxième réponse si nous savons que la personne a pour prénom Jérôme.
- la localisation des données : une information prise dans la page personnelle d'une personne a plus de chance d'être la bonne réponse.

Malheureusement, en raison d'un manque de temps, je n'ai pas pu me consacrer au calcul de cette valeur ni à sa propagation. Effectivement, lors du calcul il aurait fallu prendre en compte la valeur de confiance des données dont le champ dépend.

Par exemple si nous cherchons le téléphone d'un individu mais que nous sommes pas sûr de l'écriture exacte de son nom, alors ce doute doit être pris en compte lors du calcul d'une valeur de certitude pour le numéro de téléphone en attribuant une valeur de certitude au nom. Nous pourrions très bien imaginer donner une valeur au champ d'après le résultat du calcul suivant : (Valeur de Confiance pour le nom * Valeur de confiance pour la localité)/ le nombre de réponses trouvées.

4.4 Réalisation

Mon application devait donc regrouper toutes ces données pour pouvoir trouver de l'information. L'interface graphique n'est qu'un moyen de percevoir les données de façon structurée mais elle ne constituait en rien un but premier. Le plus important était surtout de récolter de l'information et ensuite l'analyser pour ne conserver que ce qui nous importait. Ce travail a donné lieu à plusieurs programmes d'expérimentation. En effet, j'ai séparé le travail en plusieurs étapes en essayant dans un premier temps de récupérer les pages internet, puis de tranformer la page obtenue en arbre DOM, enfin d'accéder à l'information en parcourant l'arbre. Un dernier travail a été d'insérer ces données dans mon interface graphique.

Dans la suite de ce chapitre, je présenterai les fonctionnalités de mon application consacrées à la recherche et j'expliquerai ensuite les techniques de wrappers qui permettent de trouver de l'information sur le Web quand les données initiales le permettent.

4.4.1 Action de recherche

Chaque champ de l'application devra à terme posséder les boutons GET, EXPLAIN et VALUES. Je n'ai affiché que ceux pour lesquels la recherche aboutit.

Le bouton GET permet d'obtenir de l'information en cherchant sur internet.

Le bouton EXPLAIN permet d'obtenir comment on a obtenu l'information en expliquant les dépendances entre attributs, l'URL où l'on va chercher l'information ainsi que le chemin pour extraire les éléments de connaissance

et enfin on explique comment calculer une valeur de confiance pour le champ concerné.

Le bouton VALUES donne la valeur de confiance, malheureusement comme je n'ai pas eu le temps de le faire, elle indique seulement le nombre de réponses trouvées.

En plus des boutons concernant un champ, il existe des boutons plus généraux. Je n'ai eu le temps d'en implanter qu'un seul qui est le bouton FILL.

Le bouton FILL permet de remplir tous les champs vides pour lesquels le bouton GET a été implémenté.

Trois boutons pourront être implémentés dans le futur d'après le schéma vu en chapitre 2. Ces boutons sont PUBLICATIONS, COLLABORATORS et CHECK. PUBLICATIONS doit permettre d'accéder aux publications de la personne, COLLABORATORS doit permettre d'accéder aux collaborateurs de la personne étudiée. Et enfin le bouton CHECK pourra vérifier si les champs remplis sont bien cohérents entre eux.

4.4.2 Implémentation

Pour pouvoir extraire une information précise nous nous sommes servis des techniques de wrapper. Par exemple google ou d'autres sites retournent des informations au format HTML. Il faut analyser, dans ce cas, l'information reçue (par exemple regarder ce qui est en gras). Le problème de cette technique est que nous sommes dépendants du site où l'on récupère les informations. Par exemple, si les pages jaunes changent leur cgi ou leur disposition de pages réponses alors ça ne marche plus.

Lorsque nous analysons l'arbre DOM, il faut tout mettre en minuscules et enlever les espaces non désirables pour pouvoir faire des comparaisons de chaînes de caractères intéressantes.

Les étapes allant de la connection à l'URL jusqu'a la transformation en arbre DOM sont toujours les mêmes, j'ai donc créé une classe abstraite en JAVA définissant des méthodes permettant de réaliser ces étapes. Une de ces méthodes est abstraite et doit être redéfinie dans les sous-classes. Cette fonction permet de parcourir l'arbre DOM et de trouver grâce à un chemin spécifié l'information recherchée. Une sous-classe représente l'action que le bouton GET doit effectuer et elle précise à sa classe mère l'URL à utiliser ainsi que le chemin d'accés à l'information.

Pour pouvoir trouver des éléments de connaissance dans l'arbre DOM, nous aurions pû nous servir d'une bibliothèque xPath pour JAVA. Il m'a semblé plus facile et surtout plus rapide de gérer mon propre path. Il ne faut

pas oublier de mettre à jour ce path tout au long du parcours de l'arbre DOM c'est-à-dire lors du début ou la fin d'un élément.

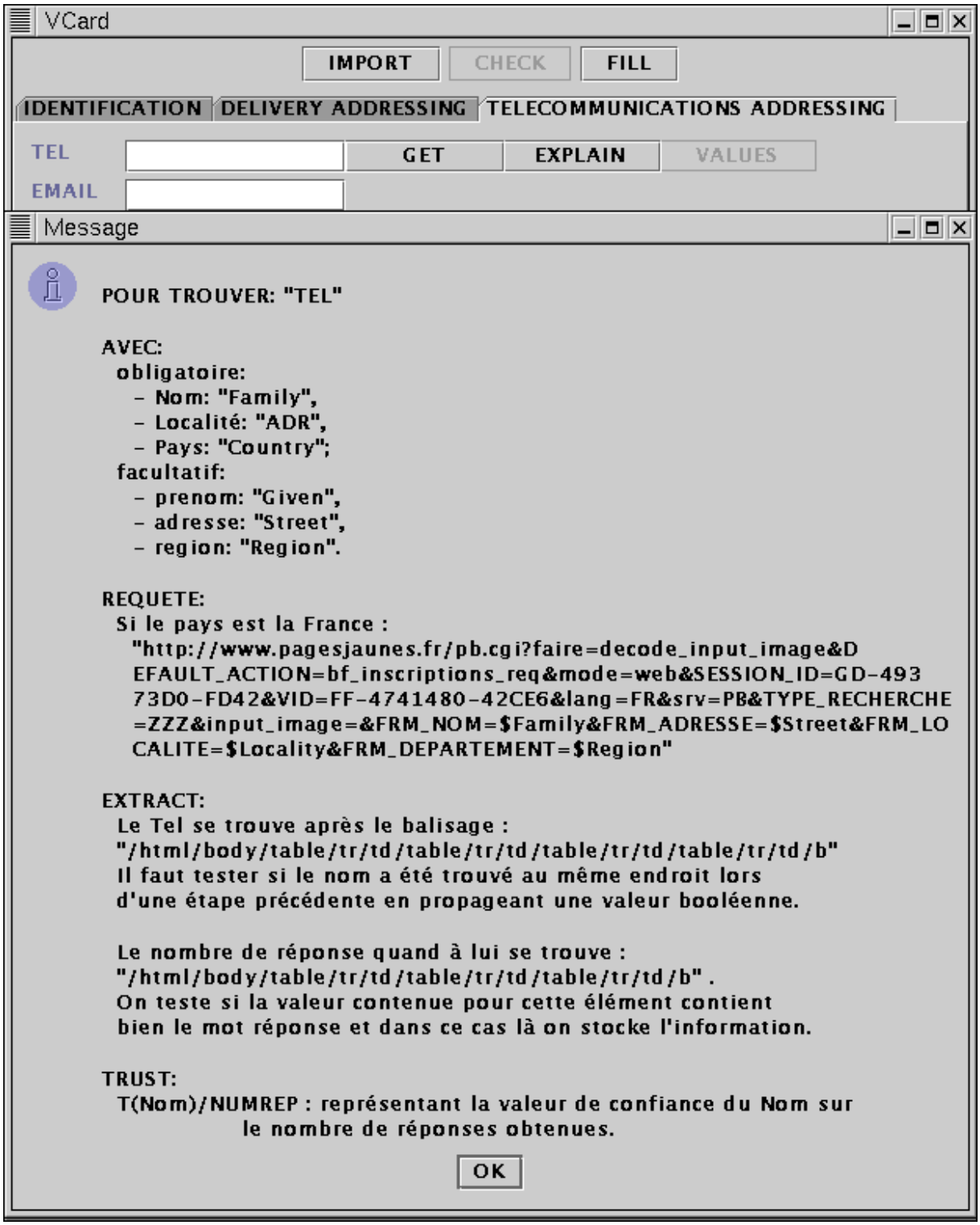

FIG. 4.1 – Appui sur le bouton Explain

FIG. 4.2 – Appui sur le bouton VALUES

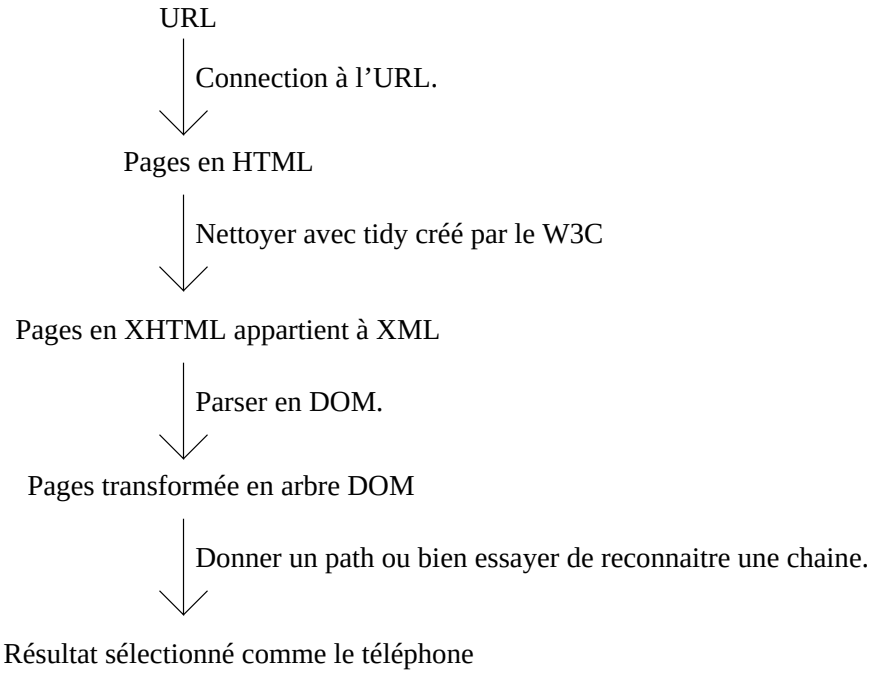

FIG. 4.3 – Description des techniques de wrapper.

Conclusion

Conclusion scientifique sur le stage

L'objectif initial étant de voir jusqu'où nous pouvions aller dans la recherche de l'information personnelle, je commenterai les résultats obtenus ainsi que les principales difficultés rencontrées puis je préciserai ce qu'il reste à faire.

Bilan

La première partie de mon projet qui consistait à intégrer une bibliothèque d'importation, d'exportation du format vCard/vCal a été terminée mi-juillet. Quelques changements de l'API ont été nécéssaires en fin de stage pour pouvoir s'occuper des champs composés. J'ai eu de gros problèmes de décodage et d'encodage avec l'API JavaMail qui n'encodait pas de la même façon qu'il décodait. Ce problème persiste encore lors de l'encodage des données.

La deuxième partie qui consistait à développer une application permettant de rechercher de l'information complétant un ou plusieurs champs de vCard et d'en donner une mesure n'est pas tout à fait terminée. Mon application trouve le téléphone, l'adresse ou encore le code postal pour les personnes habitant en France et les insère dans le format vCard. J'ai perdu beaucoup de temps à tenter de récupérer une page de résultat google pour trouver une URL d'une vCard. Et je pense que google ne permet pas aux applications Java de récupérer de l'information. Cependant, j'espère bien remplir le champ URL d'ici la fin de mon stage en changeant de moteur de recherche. Par contre, le calcul d'une mesure de confiance ne sera pas élaboré par manque de temps mais quelques pistes ont été fournies dans ce rapport.

Perspectives

Il serait intéressant d'implanter le bouton CHECK pour vérifier l'information, et mettre en relation différents formats de PIM en implantant PUBLICATIONS et COLLABORATORS. Il serait intéressant également de faire de la recherche dans une base de vCard.

Conclusion sur l'apport du stage

Le plus important pour moi était de m'insérer dans une équipe, dans un projet en IA et d'apporter à l'entreprise ainsi que d'apprendre énormément.

Entreprise

Les bibliothèques d'importation et d'exportation que j'ai créées ont permis d'ajouter des fonctionnalités au produit Transmorpher développé par EXMO.

Personnel

Le stage a été en tout point bénifique car il m'a permis d'apprendre énormément sur plusieurs plans. D'une part, il m'a permis d'approfondir les connaissances acquises tout au long de l'année notamment en XML et dans la représentation des connaissances qui, il faut bien l'avouer n'était pas une de mes matières favorites avant le déroulement du stage. Ce stage m'a permis aussi de découvrir des outils plus généraux sur l'informatique tel que ant, CVS ... Enfin, j'ai pu découvrir comment était organisé un institut de recherche tel que l'INRIA et comment je pouvais m'insérer dans une équipe comme EXMO.

En effet, j'ai découvert ce qu'était le web sémantique ainsi que les outils nécessaires tels que le RDF dont je ne me suis finalement pas servi. Ce stage m'a permis d'approfondir la connaissance que j'avais du langage JAVA ainsi que la connaissance de XSL. J'ai découvert de nombreux outils permettant de parser du XML, des bibliothèques entièrement consacrées au langage XML. Les bases du langages XML apprises pendant le DESS m'ont permis de me mettre assez rapidement en action. Ce stage m'a donné envie de continuer dans le milieu XML.

De plus, cette expérience m'a aussi donné l'occasion de me servir d'outils bien pratiques tel que ant qui permet de compiler des sources java grâce à des fichiers écrits en XML. J'ai aussi découvert LaTeX avec lequel j'ai rédigé mon rapport de stage.

Enfin ce stage m'a permis de découvrir ce qu'était la recherche et comment était organisé un laboratoire de recherche. J'ai dû aussi m'intégrer et m'organiser pour pouvoir réaliser mon travail du mieux possible. Il me reste néanmoins quelques lacunes notamment sur l'évaluation personnelle du temps de travail. Il me semble que c'est l'un des exercices les plus difficiles de l'informatique et que l'amélioration sur ce point viendra avec l'expérience.

Annexes

Annexe A

The Semantic Web lifts off

Tim Berners-Lee, *Director of W3C*, Eric Miller, *W3C Semantic Web Activity Lead* [1]

Voice browsing, Scalable Vector Graphics, Web Services, and the Semantic Web are but a few of the W3C Activities attracting media attention. This article focuses on the W3C's Semantic Web Activity and recent developments in the Semantic Web community. Although it is difficult to predict the impact of such a far-reaching technology, current implementation and signs of adoption are encouraging and developments in future research areas are extremely promising.

What is the Semantic Web ?

The Semantic Web is an extension of the current Web in which information is given well-defined meaning, enabling computers and people to work in better cooperation. The W3C Semantic Web Activity, in collaboration with a large number of researchers and industrial partners, is tasked with defining standards and technologies that allow data on the Web to be defined and linked in a way that it can be used for more effective discovery, automation, integration, and reuse across applications. The Web will reach its full potential when it becomes a environment where data can be shared and processed by automated tools as well as by people.

How might this be useful ? Suppose you want to compare the price and choice of flower bulbs that grow best in your zip code, or you want to search online catalogs from different manufactures for equivalent replacement parts for a Volvo 740. The raw information that may answer these questions, may

[1] Explication du web sémantique par ses deux principaux protagonistes.

indeed be on the Web, but it's not in a machine-usable form. You still need a person to discern meaning of the information and its relevance to your needs.

The Semantic Web addresses this problem in two ways. First, it will enable communities to expose their data so that a program doesn't have to strip the formatting, pictures and ads from a Web page to guess at the relevant bits of information. Secondly, it will allow people to write (or generate) files which explain – to a machine – the relationship between different sets of data. For example, one will be able to make a "semantic link" between a database with a "zip-code" column and a form with a "zip" field that they actually mean the same thing. This will allow machines to follow links and facilitate the integration of data from many different sources.

This notion of being able to semantically link various resources (documents, images, people, concepts, etc.) is an important one. With this we can begin to move from the current Web of simple hyperlinks to a more expressive semantically-rich Web. A Web where we can incrementally add meaning and express a whole new set of relationships (hasLocation, worksFor, isAuthorOf, hasSubjectOf, dependsOn, etc.) among resource, making explicit the particular contextual relationships that is implicit in the current Web. Thus opening new doors for effective information integration, management and automated services.

How is it being developed ?

There are two places to look for Semantic Web progress : from the ground up, in the infrastructural and architectural work coordinated by the W3C, and from top down, in application-specific work by those leveraging Semantic Web technologies in various demonstrations, applications and products. This article provides a introduction to both views with a specific focus on those areas in which the W3C is directly involved.

Enabling Standards

Uniform Resource Identifiers (URIs) are a fundamental component of the current Web and are a foundation of the Semantic Web.The Extensible Markup Language (XML) is also a fundamental component for supporting the Semantic Web. XML provides an interoperable syntactical foundation upon which the more important issue of representing relationships and meaning can be built. URIs provide the ability for uniquely identifying resources as

well as relationships among resources. The Resource Description Framework (RDF) family of standards leverages URIs and XML to provide an stepwise set of functionality to represent these relationships and meaning.

The W3C Semantic Web Activity's charter is to serve a leadership role in the design of specifications and the open, collaborative development of technologies focused on representing relationships and meaning. The base level of the RDF family of standards is a W3C Recommendation. The RDF Core Working Group is in the process of formalizing the original RDF Model and Syntax Recommendation which provides a simple yet powerful framework for representing information in the Web. Building on this work, the group is additionally defining a simple means for declaring RDF Vocabularies. RDF Vocabularies are descriptive terms (e.g. Service, Book, Image, title, description, rights, etc.) that are useful to communities recoding information in a way that enables effective reuse, integration and aggregation of data. Additional deliverables include a precise semantic theory of these standards that will support future work, as well as a primer designed to provide the reader a basic understanding of RDF and its application.

Simple data integration, aggregation and interoperability are enabled by these base level RDF standards. An increasing need for interoperability at more expressive descriptive level is also desired. The Web Ontology Working Group is charted to build upon the RDF Core work a language for defining structured, Web-based ontologies. Ontologies can be used by automated tools to power advanced services such as more accurate Web search, intelligent software agents and knowledge management. Web portals, corporate Web site management, intelligent agents and ubiquitous computing are just some of the identified scenarios that helped shaped the requirements for this work.

Advanced Development

Just as the early development of the Web depended on code modules such as libwww, W3C is devoting resources to the creation and distribution of similar core components that will form the basis for the Semantic Web. The W3C Semantic Web Advanced Development (SWAD) initiatives are designed to work in collaboration with a large number of researchers and industrial partners and stimulate complementary areas of development that will help facilitate the deployment and and future standards work associated with the Semantic Web.

SWAD DAML

The purpose of the SWAD DAML project is to contribute to the development of a vibrant, ubiquitous Semantic Web by building critical Semantic Web infrastructure and demonstrating how that infrastructure can be used by working, user-oriented applications.

SWAD DAML is designed to build on the DARPA Agent Markup Language (DAML) infrastructure to provide an interchange between two or more different applications. The first involves structured information manipulation required to maintain the ongoing activities of an organization such as the W3C, these include : access control, collaborative development, and meeting management. The second application is focused on the informal and often heuristic processes involved in document management in a personalized information environment. Integrated into both environments will be tools to enable authors to control terms under which personal or sensitive information is used by others, a critical feature to encourage sharing of semantic content.

SWAD Europe

SWAD-Europe will highlight practical examples of where real value can be added to the Web through Semantic Web technologies. The focus on this initiative is on providing practical demonstrations of how the Semantic Web can address problems in areas such as : sitemaps, news channel syndication, thesauri, classification, topic maps, calendaring, scheduling, collaboration, annotations, quality ratings, shared bookmarks, Dublin Core for simple resource discovery, web service description and discovery, trust and rights management and how to effectively and efficiently integrate these technologies together.

SWAD-Europe will additionally concentrate on exploratory implementation and pre-consensus design in areas such as querying, and the integration of multiple Semantic Web technologies. It shall provide valuable input and experiences to future standards work.

SWAD Simile

W3C is additionally working with HP, MIT Libraries, and MIT's Lab for Computer Science on Simile, which seeks to enhance interoperability among digital assets, schemas, metadata, and services across distributed individual,

community, and institutional stores and across value chains that provide useful end-user services by drawing upon the assets, schemas, and metadata held in such stores. Simile will leverage and extend DSpace, enhancing its support for arbitrary schemas and metadata, primarily though the application of RDF and semantic Web techniques. The project also aims to implement a digital asset dissemination architecture based upon Web standards, enabling services to operate upon relevant assets, schemas, and metadata within distributed stores.

The Simile effort will be grounded by focusing on well-defined, real-world use cases in the libraries' domain. Since parallel work is underway to deploy DSpace at a number of leading research libraries, we hope that such an approach will lead to a powerful deployment channel through which the utility and readiness of Semantic Web tools and techniques can be compellingly demonstrated in a visible and global community.

SWAD Oxygen

The MIT/LCS Oxygen project is designed to enable pervasive, human-centered computing through a combination of specific user and system technologies. Oxygen's user technologies directly address human needs. Speech and vision technologies enable us to communicate with Oxygen as if we're interacting with another person, saving much time and effort. Automation, individualized knowledge access, and collaboration technologies help us perform a wide variety of tasks what we want to do in the ways we like to do them. In Oxygen, these technologies enable the formation of spontaneous collaborative regions that provide support for recording, archiving, and linking fragments of meeting records to issues, summaries, keywords, and annotations.

The Semantic Web is designed to foster similar collaborative environment and the W3C is working with project Oxygen to help support this goal. The ability for "anyone to say anything about anything" is an important characteristic of the current Web and is a fundamental principal of the Semantic Web. Knowing who is making these assertions is increasingly important in trusting these descriptions and enabling a 'Web of Trust'. The Annotea advanced development project provides the basis for asserting descriptive information, comments, notes, reviews, explanations, or other types of external remarks to any resource. Together with XML digital signatures, the Annotea project will provide a test-bed for 'Web-of-Trust' Semantic Web applications.

Applications - spinning upward

Though not the focus of this article, the deployment of RDF based technologies is increasingly significant. The W3C Semantic Web Activity hosts the RDF Interest Group which coordinates public implementation, and shares deployment experiences of these technologies. Arising out of RDF Interest Group discussions are several public issue-specific mailing lists, including RDF-based calendar and group scheduling systems, logic-based languages, queries and rules for RDF data and distributed annotation and collaboration systems. Each of these discussion groups are designed to focus on complementary areas of interest associated with the Semantic Web Activity. Each of which fosters cooperation and collaboration among individuals and organization working on related Semantic Web technologies.

In addition to these Interest Group lists there are a variety of domain specific communities who are using RDF/XML to publish their data on the Web. These notably include the Dublin Core Metadata Initiative, the IMS Global Learning Consortium vocabularies for facilitating online distributed learning, XMLnews, PRISM, the RDF Site Summary (RSS 1.0) for supporting news syndication, Musicbrainz for cataloging and cross referencing music, and Creative Commons for supporting digital rights description to name but a few. The Topic Map (XTM) community has been finding increasing synergy with the RDF data model.

Early commercial adopters such as Adobe's eXtensible Metadata Platform (XMP), for example, leverage RDF/XML to enable more effective management of digital resources. Adobe applications and workflow partners through XMP can leverage the power of RDF/XML to provide a standardized means for supporting the creation, processing, and interchange of document metadata across publishing workflows. This in-turn reduces cost and makes for more effective management of digital resources possible both within and across organizational boundaries.

New things opening up

The most exciting thing about the Semantic Web is not what we can imagine doing with it, but what we can't yet imagine it will do. Just as global indexes, and Google's algorithms were not dreamed of in the early web days, we cannot imagine now all the new research challenges and exciting product

areas which will appear once there is a web of data to explore. Many existing fields for knowledge representation and data management have typically made assumptions regarding a conceptually or physically centralized system, and as such their application to the Semantic Web is not straightforward. Given a mass of rules relating data in different vocabularies, and an unbounded set of datasets in different vocabularies, what algorithms will efficiently resolve general queries? What conventions for the storage of tips and pointers will allow data to be reused and converted automatically? What techniques will allow a system to operate securely while processing very diverse data from untrusted agents? How can one represent – and then implement – personal privacy in such a world?

The Semantic Web starts as a simple circles-and-arrows diagram relating things, which slowly expands and coalesces to become global and vast. The WWW of human-readable documents spawned a social revolution. The Semantic Web may in turn spawn a revolution in computing. In neither case a change in the power of one person or one computer, but in each case a dramatic change in the role it can play in the world, by being able to find out almost anything it needs to know virtually immediately.

Annexe B

Champs des vCard

Champs d'une vCard	Description pour l'objet représenté par la vCard	Version
FN	Le nom formatté.	obligatoire dans la version 3.0
N	Le nom structuré. Les champs sont séparés par des points virgules : – Family : le nom de famille ; – Given : le prénom ; – Other : les prénoms complémentaires séparés par des virgules ; – Prefix : les préfixes honorifiques séparés par des virgules ; – Suffix : les suffixes honorifiques séparés par des virgules.	obligatoire dans la version 3.0
NICKNAME	Le surnom.	nouveauté dans la version 3.0
PHOTO	Une image ou une photographie qui montre quelques aspects de l'objet représenté par une vCard.	
BDAY	la date de naissance.	

Champs d'une vCard	Description pour l'objet représenté par la vCard	Version
ADR	L'adresse structurée. Les champs sont séparés par des points virgules : – Pobox : la boîte postale ; – Extadd : l'extension d'adresse ; – Street : l'adresse ; – Locality : la localité ; – Region : la région ; – Pcode : le code postal ; – Country : le pays.	
LABEL	L'adresse formattée.	
TEL	Le numéro de téléphone.	
EMAIL	L'adresse électronique.	
MAILER	Le logiciel utilisé pour s'occuper des mails.	
TZ	Les informations concernant les fuseaux horaires.	
GEO	Position sur le globe	
TITLE	Titre du travail.	
ROLE	Le rôle, l'occupation ou la catégorie sociale.	
LOGO	Un logo	
AGENT	Des informatons à propos d'une personne qui joue un rôle.	
ORG	Le nom et les unités – Orgname : le nom de l'organisation ; – Orgunit : l'unité de l'organisation.	
CATEGORIES	Information sur la catégorie sociale.	nouveauté dans la version 3.0
NOTE	Informations supplémentaires ou commentaires.	
PRODID	La marque du produit qui a créé la vCard	nouveauté dans la version 3.0
REV	Dates de dernière révision.	

Champs d'une vCard	Description pour l'objet représenté par la vCard	Version
SORT-STRING	Définit la partie du nom qui est importante.	nouveauté dans la version 3.0
SOUND	Un son qui représente quelques aspects de la vCard tel que la prononciation du nom.	
UID	Un identifiant unique.	
URL	Une URL.	
VERSION	La version utilisée.	obligatoire dans la version 3.0
CLASS	La classification d'accés.	nouveauté dans la version 3.0
KEY	Une clé publique ou un certificat d'identification.	
X-word	Le vocabulaire peut être étendu avec des champs privé commençant par X- .	

Annexe C

Vocabulaire FOAF

Il y a quatre classes principales qui ont chacune des propriétés : Organization, Project, Person et Document.

Dans la suite de cette annexe, je vais vous présenter les propriétés.

Nom de la propriété	Description
page	Une page ou un document concernant la classe décrite.
schoolHomepage	Une page d'accueil de l'école de la personne.
surname	Le nom de famille d'une Personne.
id	Un identifiant pour la classe décrite.
knows	Une personne connue par cette personne.
title	Titre (M., Mme, Mlle., Dr. etc..)
mbox	Une boîte aux lettres électronique est associée à exactement un propriétaire.
theme	Thème
interest	Une page au sujet d'un intérêt pour cette personne
phone	Un numéro de téléphone international entièrement qualifié, indiqué à l'aide de : (refs : http://www.w3.org/Addressing/schemes.html#tel)
publications	Un lien vers les publications de cette personne

Nom de la propriété	Description
nick	Un surnom caractérisant un agent.
currentProject	Un projet en cours pour cette personne.
name	Un nom pour la classe décrite.
pastProject	Un projet sur lequel la personne a travaillé précédemment.
firstName	Le prénom d'une personne
fundedBy	
homepage	Une page d'accueil pour la classe décrite
mbox_sha1sum	
linkedWith	Lien générique
img	Une image qui peut être employée pour représenter la classe décrite
geekcode	
depiction	Une description de la classe décrite
logo	Un logo représentant la classe décrite
workplaceHomepage	Une page d'accueil du lieu de travail d'une personne ; la page d'accueil d'une organisation pour laquelle il travaille.
dnaChecksum	
topic	Le sujet d'une page ou d'un document
workInfoHomepage	Une page d'accueil donnant de l'information sur le travail d'une personne ; une page au sujet du travail d'une organisation.
givenname	Le prénom donné à un agent.

Annexe D

Le format des fichiers bib

Le tableau suivant décrit le rôle des champs.

Nom du champ	Rôle
address	La ville ou l'adresse complète de l'éditeur.
author	Les noms des auteurs.
booktitle	Le titre du livre utilisé dans le cas où l'on cite une partie du livre.
chapter	Le chapitre que l'on cite dans un livre.
crossref	Une référence croisée.
edition	Indique le numéro de l'édition, ou plutôt son ordinal.
editor	Indique le nom du rédacteur en chef.
howpublished	Ne sert que dans le cas où le document cité n'est pas d'un type classique, comme un livre, un article de journal ou de conférence, etc.
institution	Dans le cas d'un rapport, indique l'institution qui l'a publié.
journal	Le nom du journal dont est extrait un article.
key	Permet de définir le label d'une entrée, dans le cas où il ne peut pas être calculé par BibTEX.
month	Le mois de parution du document, s'il est connu.
note	Une remarque additionnelle quelconque.
number	Un numéro : le numéro dans une série ou le numéro d'un rapport.
organization	L'organisateur d'une conférence.
pages	Les pages qui nous intéressent.
publisher	Pour le nom de la maison d'édition ou de l'organisme qui a publié le document cité.

Nom du champ	Rôle
school	Pour un mémoire ou une thèse, l'école où il a été réalisé.
series	Le nom d'une série d'ouvrages, d'une collection de bouquins, ...
title	Le titre du document que l'on cite.
type	Le type de publication, au cas où ce n'est pas clair.
volume	Le numéro de volume dans une série ou dans une collection.
year	L'année de publication du document cité.

Le tableau suivant décrit les différents types d'entrée.

Type d'entrée	Utilisation
@article	Article paru dans un journal.
@book	Un livre.
@booklet	Un *petit* livre, sans champ publisher.
@conférence	Article dans les actes d'une conférence, d'un colloque, d'une rencontre...
@inbook	Partie (un chapitre, souvent) d'un livre.
@incollection	Grosse partie d'un livre. Pas juste un petit chapitre... Il faut en particulier que cette partie ait son propre titre.
@inproceedings	Exactement pareil que **@conference**.
@manual	Un manuel, une petite doc.
@masterthesis	Mémoire de D.E.A., ou équivalent.
@misc	Quand on ne sait pas quoi mettre, on met @misc...
@phdthesis	Thèse de doctorat, d'habilitation, ou un autre gros truc dans le même genre.
@proceedings	Actes d'une conférence.
@techreport	Rapport technique, publié par un labo, un centre de recherche, ...
@unpublished	Un document non publié. C'est souvent assez proche de l'entrée @misc, sauf que là, il y a un auteur et un titre.

Annexe E

Eléments de métadonnées du Dublin Core, Version 1.1 : Description de Réference

Cette annexe résume les définitions des éléments de métadonnées du Dublin Core définis dans le document[RFC2413].

Elément	Définition
Title	Le nom donné à la ressource.
Creator	L'entité principalement responsable de la création du contenu de la ressource.
Subject	Le sujet du contenu de la ressource.
Description	Une description du contenu de la ressource.
Publisher	L'entité responsable de la diffusion de la ressource, dans sa forme actuelle, comme un département universitaire, une entreprise.
Contributor	Une entité qui a contribué à la création du contenu de la ressource.
Date	Une date associée avec un événement dans le cycle de vie de la ressource.
Type	La nature ou le genre du contenu de la ressource.
Format	La matérialisation physique ou digitale de la ressource.
Identifier	Une référence non ambiguë à la ressource dans un contexte donné.
Source	Une référence à une ressource à partir de laquelle la ressource actuelle a été dérivée.
Language	La langue du contenu intellectuel de la ressource.
Relation	Une référence à une autre ressource qui a un rapport avec cette ressource.
Coverage	La portée ou la couverture spatio-temporelle de la ressource.
Rights	Information sur les droits sur et au sujet de la ressource.

Bibliographie

[1] Dan Brickley, Libby Miller, and rdfweb-dev listmembers. FOAF : the friend of a friend vocabulary. Technical report, RDFWeb, 2002. http ://xmlns.com/foaf/0.1/.

[2] F. Dawson and T. Howes. vCard MIME Directory Profile. Technical report, Network Working Group, 1998. http ://www.faqs.org/rfcs/rfc2426.html.

[3] F. Dawson and D. Stenerson. Internet Calendaring and Scheduling Core Object Specification. Technical report, Network Working Group, 1998. http ://www.faqs.org/rfcs/rfc2445.html.

[4] Nicolas Markey. *Tame the BeaST BibTeX, de B à X...*.

[5] Eric M.Burke. *Java et XSLT*. O'REILLY, 2001.

[6] Brett McLaughlin. *Java & XML*. O'REILLY, deuxième edition, 2001.

[7] Iannella Renato. Representing vCard Objects in RDF/XML. Technical report, W3C, 1998. http ://www.w3.org/TR/vcard-rdf.

[8] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. Technical report, Network Working Group, 1998. http ://www.faqs.org/rfcs/rfc2413.html.